TESTING FOR STRUCTURAL BREAKS IN THE EVALUATION OF PROGRAMS

Anne Morrison Piehl, Suzanne J. Cooper, Anthony A. Braga, and David M. Kennedy*

Abstract—A youth homicide reduction initiative in Boston in the mid-1990s poses particular difficulties for program evaluation because it did not have a control group and the exact implementation date is unknown. A standard methodology in program evaluation is to use time series variation to compare pre- and postprogram outcomes. Such an approach is not valid, however, when the timing of program implementation or effect is unknown. To evaluate the Boston initiative, we adapt from the time series literature an unknown-breakpoint test to test for a change in regime. Tests for parameter instability provide a flexible framework for testing a range of hypotheses commonly posed in program evaluation. These tests both pinpoint the timing of maximal break and provide a valid test of statistical significance. We evaluate the results of the estimation using the asymptotic results in the literature and with our own Monte Carlo analyses. We conclude there was a statistically significant discontinuity in youth homicide incidents (on the order of 60%) shortly after the intervention was unveiled.

I. Introduction

An initiative to reduce youth homicide citywide was unveiled in Boston in the summer of 1996. The pre-post comparison is impressive: the number of young victims of homicide fell from 3.8 per month in 1995 to 1.25 in 1997. Yet a scientifically valid evaluation requires more than the test of difference in a simple time series. The usual approach in such settings is to compare the difference in differences. But in this example, because of the design of the program, the precise implementation date is unknown. Perhaps more important, there is no appropriate control or even comparison group, as all those people and locations at high risk were targeted by the program. As a result, neither difference can be properly computed.

Given that a difference-in-differences research design is not available, we turn to the time series literature on unknown breakpoints for guidance. This literature has concerned itself with identifying regime changes using time series data, which seems appropriate for evaluating an initiative that aimed to change the regime under which youth behaved on the streets of Boston. Implementation lags in policy interventions make it very difficult to say that a particular date defines the break between the preprogram and postprogram periods, even when the implementation date is known. In such settings, if a researcher were to use the time series to judge where to time the break, even informally, the test would not have the correct critical values. Rather, statistical significance in a traditional pre-post analysis is overstated if ex post information is used to locate the time when the program is said to have begun.

In this paper we build on the literature on unknown breakpoints, applying it in a new setting, program evaluation. This setup allows us to report new results that (1) test in the most general way for the existence of a break in the data series and (2) pinpoint the timing of the break. From these results we can calculate the magnitude of the impact. We use the methods to evaluate the Boston Gun Project and the intervention it generated, Operation Ceasefire. In addition, we explore alternative counterfactuals in order to assess the validity of the interpretation of the break as a program effect. Since the existing time series literature provides only asymptotic distributions for the statistics used, we report results of Monte Carlo analyses to give us some insight into the use of this technique with finite samples. We conclude with a discussion of the ways in which the technique can be applied to other program evaluation cases with possibly different characteristics.

II. The Boston Project

A. History

In the early 1990s, amid concern about the growing numbers of youth involved in homicides, both as victims and as offenders, a multiagency working group was convened to research Boston's youth violence problem, to craft a strategy to respond to the conditions, to implement that strategy, and to evaluate the experience.[1] Throughout the effort, the goal of the Boston Gun Project (BGP) was to reduce youth homicide in the city of Boston in the near term. Given that the agencies represented in the working group had responsibility for the city as a whole, it was not possible to designate a priori a portion of the city as a control site.

Research undertaken by the group indicated that youth homicides were concentrated geographically (most incidents occurred within the three neighborhoods of Roxbury, Dorchester, and Mattapan) and demographically (most victims were male and African American). Furthermore, an incident-by-incident analysis revealed that, conservatively, a minimum of 60% of the homicides could be characterized

Received for publication September 14, 2000. Revision accepted for publication September 20, 2002.

* John F. Kennedy School of Government, Harvard University.

The Boston Gun Project was supported under award #904-IJ-CX-0056 from the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. Piehl appreciates the support of the Robert Wood Johnson Foundation. Points of view in this document are those of the authors and do not necessarily represent the official position of the U.S. Department of Justice or the Foundation. The authors appreciate the helpful comments of John Campbell and two anonymous referees, as well as Christopher Avery, Richard Blundell, Kristin Butcher, David Card, Nolan Miller, James Stock, and seminar participants at the University of California at Berkeley, Harvard University, Yale University, and the NBER.

[1] Membership in the working group included representatives from the Boston Police Department; local and federal prosecutors; the Bureau of Alcohol, Tobacco and Firearms; probation and parole departments; community outreach workers; and academics.

The Review of Economics and Statistics, August 2003, 85(3): 550–558

as arising out of a nexus of disputes across gangs.[2] These gangs were involved in longstanding antagonisms with each other, which sometimes erupted in series of retaliations. These disputes were not necessarily confined geographically, a second reason why it was not possible to designate control sites for the intervention.[3]

The majority of the youth homicide incidents were even further concentrated among a group of young men with whom many in law enforcement were familiar due to criminal activity and gang associations.[4] The working group designed a strategy to target deterrence messages and sanctions toward those individuals and groups most active in perpetrating violence, and to do so ex ante. This was a departure from the usual practice of concentrating enforcement efforts on investigation after crimes occurred. From the point of view of prevention, there were two key premises for the strategy. First, *key personnel* were identifiable. Second, retaliations were responsible for an appreciable amount of the violence. If key gangs and key individuals within those gangs could be identified and reached by a deterrence message, the intergang dynamic could be interrupted. As is true in the general tipping model, if the interruption of the dynamic were substantial enough, there could be a dramatic shift in the level of violence. In short, the regime on the streets could be moved from a high-violence equilibrium to a low-violence equilibrium.[5]

Based on these premises, an intervention known as Operation Ceasefire was developed. Because law enforcement resources in the city were limited, choices had to be made as to which incidents and threats could be addressed. One element of the intervention involved members of the working group sharing information to identify those involved in violent disputes in order to target resources. Enforcement agencies customized sanctions to individuals and gangs depending on their (often extensive) prior involvement with the criminal justice system by enforcing conditions of probation and/or parole. Due to the longstanding nature of antagonisms between gangs, practitioners could predict conflicts with some success. In such cases, both enforcement and social services mobilized to prevent retaliations. Finally, in order to deter initiations of violence and retaliations, the working group advertised the goals, capacities, and achievements of the initiative to individuals and gangs identified as at risk of violent assault or victimization.

The project evolved as follows. The working group was constituted early in 1995 and met regularly during that calendar year to research the basic problem of youth homicide. Results were presented to the Boston Police Department and other partners in January of 1996. Based on this research, the strategy to be implemented was defined and refined over the spring. In May, the first formal event was held, a forum explaining the strategy to a group of gang members. At the end of June the strategy was publicly announced to the public at a conference organized by the Boston Bar Association.[6]

B. *Evaluation Issues*

Two attributes of the intervention pose particular challenges for evaluation design. First, the project evolved over time, moving from the research phase to strategy development and then to implementation. Even once the strategy was in place, the intervention's attributes changed with the circumstances of particular violent outbursts and as the nature of youth violence in the city evolved. As a result, it is impossible to assert with confidence the date on which implementation of the intervention began and even more difficult to determine the date of any effect without looking at the data. Therefore, an evaluation should treat the timing of any program effect as unknown. Second, the hypothesis of the intervention design was that affecting a dispute between two gangs would naturally have spillovers to other groups with which the original gangs feuded. The spillovers were expected to result from particular enforcement actions. Further, the working group hoped the deterrence message would be heard by others uninvolved in the original dispute.

As a result of these features of the Boston initiative, treatment could not be randomized over individuals. There are several reasonable alternative evaluation designs in this case. In a panel of cities, the experiences of youth in other cities provide the counterfactual for what would have been expected to occur in Boston in the absence of the program. However, trends vary greatly across cities, and in addition there may well have been spillovers as other cities adopted aspects of the program after it received positive media attention early on. Therefore it is not obvious that using other cities is preferable to using control variables within Boston.

An alternative is to use a control group within the city of Boston. However, there are two problems with using the number of adult homicide victims to provide a counterfactual for the number of youth homicide victims. First, the youth offenders targeted by the program would be expected to commit crimes against not only other youth, but also adults.[7] For this reason, the number of adult homicide

[2] Relative to gangs in other cities, Boston gangs were small with somewhat loose organization. For example, the high degree of corporatization reported by Levitt and Venkatesh (2000) was not the norm observed in Boston.

[3] Here we discuss the conceptual impossibility of utilizing control sites. It would also have been difficult or impossible in practice. It is highly unlikely that any of the agencies involved in the project would have agreed to set aside certain sections of the city as control sites for the purpose of improving the evaluation design.

[4] For more details on the basic research on youth homicide in Boston, see Kennedy, Piehl, and Braga (1996).

[5] This characterization was laid out ex ante (Kennedy et al., 1996), not as an explanation for the results that appear later in the current paper.

[6] For a detailed time line and a more discursive description of the details of the intervention, see Piehl, Kennedy, and Braga (2000).

[7] Using data from Supplemental Homicide Reports, Cook and Laub (1998, table 5) report that young homicide offenders tend to kill people who are older than they. For killers aged 13–17, 75% of their victims are

victims would be expected to fall in direct response to the program. In addition, the program would likely affect the offending of adults who were only slightly older than the youth targeted. The number of adult homicide victims would again fall as a result of the spillover of the program. For these reasons, adult homicide is not a usable control group for assessing the impact of the program targeted at youth.

Because of the problems with these approaches, in this paper we utilize the time series variation within Boston alone. We apply a technique to locate a discontinuity in youth homicide—that is, the number of youth homicide victims—while controlling for characteristics of the city that arguably changed over time in a manner consistent with the observed changes in youth violence. In the robustness section of the paper, we assess our findings relative to the trend in youth homicide in other cities and relative to changes in adult homicide in Boston.

C. Descriptive Statistics

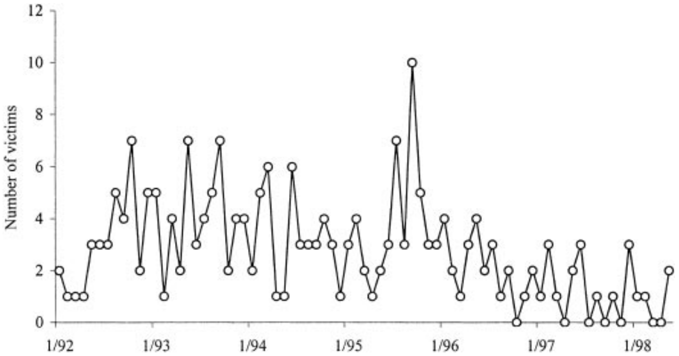
The dependent variable for the evaluation of the initiative is the monthly number of homicide victims aged 24 or under, provided by the Boston Police Department. Figure 1 plots the raw data from January 1992 through May 1998. The number of homicides is small; a number of months record zero events. The series exhibits a great deal of variation, some of which is seasonal (August and September having the highest homicide rates). This seasonal variation notwithstanding, one can see that the number of homicides is particularly low late in the time series.

Table 1 reports descriptive statistics by year for youth homicide, population, and several additional controls.[8] The first column shows that the average number of youth homicides per month was between 3 and 4 in the early 1990s. The number of incidents falls to around 1 per month by the end of the period.

In explaining changes in youth homicide, it is potentially important to control for the size of the population at risk. The second column reports the number of African American males aged 15–24, as most victims are members of this demographic group.[9] The number in this group fell by 7% over the time series. It is worth noting that the results are not sensitive to the particular population control used. In fact, the population variable is generally not statistically significant in the regressions reported below.

Table 1 includes values for several additional variables, which will be used later as controls. The number of older homicide victims also fell over this period, whether measured as all aged 25 and older (*adult* homicides) or those aged 35–44. Both of these measures are used to help determine whether the drop in youth homicide is plausibly related to the program. To control for the booming economy, we use the teen unemployment rate, which is reported in the final column. Unfortunately, due to small sample sizes in the Current Population Survey, this variable is only available annually, and only for the state as a whole. The teen unemployment rate fell by over half during this period.

FIGURE 1.—MONTHLY YOUTH HOMICIDE COUNT, 1/92–5/98



III. Tests of Structural Change

The common practice in the program evaluation literature of defining a dummy variable for the postprogram period and testing for a change in mean or other parameters can be conceptually flawed on two grounds. First, the precise timing of an effect of a program is not known, even if the start date of the program itself is known. Therefore, the postprogram dummy variable may not actually enter at the appropriate time for evaluating an effect of the program, and for this reason the estimated program effect may not be correct. Second, as Banerjee, Lumsdaine, and Stock (1992) and others point out, when the breakpoint in the data is not known a priori, conventional hypothesis testing is not valid.

In the program evaluation setting, these issues are particularly salient. Even when the start date of a program is known, implementation lags and even leads (through *announcement effects* where the program has an effect before it even begins) make it virtually impossible to identify the timing of an effect a priori. Rather, one determines when the effect occurred by looking at the data. It is this process that renders the conventional critical values for a Wald statistic invalid.

What we propose (and apply in section IV) is to use a test, developed in the time series literature, for structural change in the case of an unknown breakpoint. The branch of the literature most applicable to the program evaluation setting has its origins in the Quandt likelihood ratio (QLR) statistic (Quandt, 1960). This statistic is the maximum F-statistic from a Chow test, evaluated over all possible breakpoints. More recently, Andrews (1993) has tabulated asymptotic

older than the killer, and over 50% of victims are more than five years older than the killer.

[8] In order to avoid introducing any breaks into the data, we linearly interpolated the data that are available only annually (namely, the population and the teen unemployment rate). The results are not sensitive to the choice of month for placement of the annual value for interpolation.

[9] Of the 169 homicide victims aged 24 and under from 1992 to 1994, 72% were black males. (Authors' calculations from the Supplementary Homicide Reports.)

TABLE 1.—DESCRIPTIVE STATISTICS: MEANS (STANDARD ERRORS)

Year	Youth Homicides (24 and under)	Population, Black Males, 15–24	Adult Homicides 25 and older	Adult Homicides 35–44	Population 25–44	Teen Unemployment Rate (%)
1992	3.083 (0.543)	12,977	3.250 (0.676)	0.833 (0.297)	228,465	20.2
1993	4.000 (0.537)	12,455	4.167 (0.458)	1.083 (0.379)	227,218	18.8
1994	3.167 (0.520)	12,272	3.917 (0.802)	1.333 (0.396)	226,611	15.9
1995	3.833 (0.716)	12,222	4.167 (0.815)	1.167 (0.322)	231,367	14.7
1996	2.083 (0.358)	11,895	2.667 (0.512)	1.083 (0.288)	230,744	13.8
1997	1.250 (0.351)	12,038	2.333 (0.355)	0.750 (0.279)	228,696	12.6
1998	0.800 (0.374)	12,359	1.400 (0.748)	0.400 (0.400)	224,471	8.8

Sources: Homicide data were provided by the Boston Police Department. Population data came from the Bureau of the Census, and unemployment rates for teens in Massachusetts were provided by the Department of Employment and Training (unpublished data).

Notes: The homicide data are monthly. The population and unemployment variables are reported annually, and have been linearly interpolated. 1998 contains data only through May.

critical values for the maximum Wald statistic for the test of change in regression parameters over time, and others have extended this literature to allow tests of changes in a subset of the parameters [see Stock (1994) for a complete review].

Given a stationary time series,[10] a Wald statistic for the null hypothesis that the parameters of interest do not change between periods can be defined:

$$H_0 : \beta_t = \beta_0 \quad \text{for all } t,$$

$$H_A(\pi) : \beta_t = \begin{cases} \beta_1, & t = 1, \ldots, T\pi, \\ \beta_2, & t = T\pi + 1, \ldots, T, \end{cases}$$

where $\pi \in (0, 1)$ is the fraction of the sample before the point of parameter change, that is, $T\pi$ is the time of the change. There can in addition be another parameter vector δ_0, which is invariant with respect to π. The maximum of the Wald statistics over all possible breaks in the parameters (that is, the sup Wald statistic) is the test statistic of interest.[11] In other words, one can test the null hypothesis that the parameters do not change against the alternative hypothesis that a particular subset of the parameters does change.[12] Clearly, conventional critical values for the Wald statistic do not apply to the sup Wald statistic.

In the program evaluation context, a finding of a sup Wald statistic that does not exceed the critical value is interpreted as finding no program effect. Even when the null is rejected and a break in the parameters of a regression relationship has been found, this finding is not proof that the break was caused by the program intervention without further analysis. Of course, this limitation in interpretation is equally present in simple pre-post program analyses, and we address this concern in detail in section VI.

In the program evaluation setting, a completely agnostic search for a break anywhere in the time series, such as is often used in time series applications, is often inappropriate. Even though we may not know the exact timing of any effect, if one exists, we almost certainly have some information regarding the earliest possible date an effect could be observed, for example, the timing of initial planning for the program. Institutional knowledge, but not examination of the data, can be used to define this window of dates.[13] Certainly one is more likely to have a priori information regarding a window of dates in which a program effect could take place than to have a priori knowledge of a specific break date.

To test for a break in parameters within a prespecified window, one chooses parameters π_1 and π_2 to define the window by trimming different proportions from each end of the time series, i.e., π_1% from the beginning of the time series and $(1 - \pi_2)$% from the end.[14] The time series literature focuses mainly on symmetric trimming of data from the ends of the time series, where the trimming parameter is generally set to reserve just enough data to estimate the Wald statistics. Although asymmetric trimming has not generally been utilized in the time series literature, as the applications have not required it, this technique is particularly relevant in the program evaluation context.[15]

[10] Stationarity can be confirmed using a Dickey-Fuller test, as we do for the Boston data.

[11] Bai (1997) derives a confidence interval for the location of the breakpoint. However this is not applied here, because in the program evaluation setting the primary goal is to determine the existence of any program effect.

[12] There are, of course, many alternatives to the sup Wald test applied here. The CUSUM statistic (Brown, Durbin, and Evans, 1975), the Nyblom (1989) statistic, and the approach of Andrews and Ploberger (1994) all address the problem of the unknown breakpoint (change point) in analogous but somewhat different ways. Generally these approaches evaluate the adequacy of the model in a diagnostic sense and are not specifically geared toward identifying the location of the break, although the break can be identified. We prefer the sup Wald test here for several reasons: (a) our goal is broader than an assessment of model accuracy, (b) the alternative hypothesis is well specified, and (c) the sup Wald test has a particularly compelling intuition for the program evaluation case.

[13] Restricting one's search to a prespecified set of dates assumes there exists no change in parameters outside this window. If this assumption is violated, the sup Wald test may fail to detect a break within the window or may misestimate its magnitude.

[14] The most extreme trimming one could do is to look for a break at a single point in time. This case collapses to the Chow test.

[15] The asymptotic results for testing for a structural break with asymmetric trimming can be derived from the table in Andrews (1993). As Andrews notes, the relationship between the results for symmetric and asymmetric trimming depend on $[\pi_2^*(1 - \pi_1)]/[\pi_1^*(1 - \pi_2)]$. Given this parameter, one can interpolate the values in table 1 of Andrews (1993) to find the asymptotic critical values associated with the window for any given application.

TABLE 2.—PARAMETER INSTABILITY IN YOUTH HOMICIDE: TESTS FOR BREAK IN MEAN—VARIOUS SETS OF CONTROL VARIABLES

Model	Maximum Wald Statistic	Month of Max.	Adult Homicide Rate	Teen Unemp. Rate	Effect Size
A	33.70	June 1996	—	—	−2.49 (72%)
B	20.93	August 1996	Yes	—	−2.12 (62%)
C	8.89	June 1996	Yes	Yes	−2.02 (59%)

Sources: See Table 1 for descriptions of the variables.
Notes: 11 month indicators and population are included in all specifications in addition to the controls noted. $N = 77$ months, January 1992 through May 1998.

IV. Empirical Results

This section tests for a structural break in the Boston youth homicide time series using the sup Wald methodology. Here, we test for a change in the mean number of youth homicides. The initiative hoped to move the level of homicide to a new, lower equilibrium. That is, controlling for other factors, if there was a program effect, it should appear as a discrete shift.[16] As a robustness check, we consider alternative specifications of the form of the break.

As discussed in the previous section, we can define a range of dates for a possible treatment effect. In order to be as agnostic as possible with regard to potential announcement effects or implementation lags, we chose a wide window. We considered January 1996 through May 1997 to cover all possible dates for program impact, giving generous room for announcement effects, and 12 months following initial implementation, to allow for lagged impact.[17]

Table 2 reports the results of searching for the maximal break in the monthly number of youth homicide victims from January 1996 through May 1997 for three sets of control variables.[18] Model A includes controls for the population of black males aged 15–24 and a full set of month dummies. In this specification the sup Wald statistic (33.70) occurred in June 1996. This test statistic far exceeds the conventional critical values for a Wald statistic; for example, the 5% Chow critical value is 3.97 for 77 observations. Furthermore, the test statistic exceeds the 5% asymptotic critical value for a break in one parameter with a search window this size, which is 6.62.[19] However, because of our small sample, we do not feel comfortable relying on the asymptotic critical values and therefore, in the next section, assess statistical significance with Monte Carlo results. The effect size is reported in the final column: a reduction of 2.49 homicide victims per month, or an approximately 72% decline.

In model B, we used in addition the rate of adult homicide victimization to control for the counterfactual for youth homicide in the absence of the program. As mentioned in section II, this is quite a strict test, given that the intervention could very well have affected the victimization of older people. With adult homicide as a control, the break continues to be located in the summer of 1996 (now August), the value of the sup Wald statistic is 20.9, and the effect size is somewhat lower (a 2.12 victim reduction per month).

Model C adds the teen unemployment rate to the controls for the previous model. Adding this covariate reduced the maximal value of the Wald statistic by over half (to 8.89). Since the unemployment rate in general was falling over the time period of this analysis, these results are not surprising. The break was still placed in June 1996, and the effect size remained near 60%. At the same time, the unemployment rate was not statistically significant in explaining the number of youth homicides when a break is specified in the maximal month. The finding that unemployment is not strongly related to crime is not unusual in the literature (Piehl, 1998). The precise timing in the relationship between unemployment and youth homicide appears to be coincidental. Unemployment was falling continuously during this time due to a general macroeconomic boom. For this reason, the unemployment rate acts much like a linear trend in this analysis. We further explore the role of a linear trend in section VI.

The results of these three specifications are easily summarized by plotting the value of the Wald statistic over the period considered. From figure 2 it is clear that the three sets of controls lead to similar patterns over time. In each case, the highest value of the break statistic is in the summer of 1996, with a rather dramatic falloff on either side. Also, the lower statistical significance is revealed in the height of the lines (as it was in table 2). Interestingly, the placement of a pre-post dummy in any month in this window in specification A or B would have exceeded the Chow value. The systematic search for maximal break pins down the timing.

V. Finite-Sample Properties

As noted earlier, the asymptotic critical values may not be appropriate for determining the statistical significance of the Wald statistics reported in table 2, because our sample size

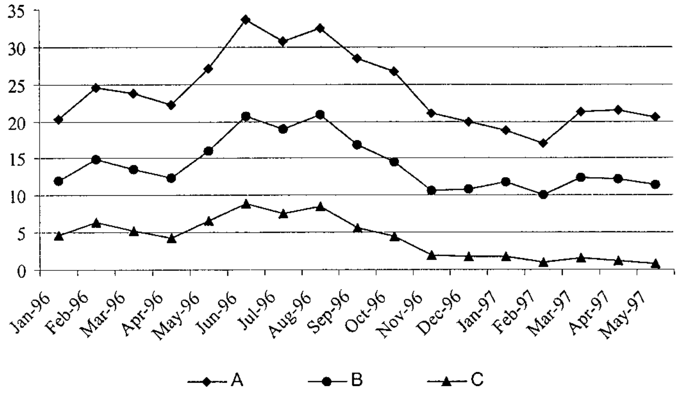
[16] There is no reason, from what is known about youth homicide, to believe that there would be changes in the seasonality or in the relationship between economic conditions, for example, and homicide.

[17] The results presented here are invariant to the choice of window.

[18] Because of the count nature of the data, there are three estimation options: Poisson, negative binomial, and OLS (with robust standard errors). In the Poisson specification, a test for a break in the mean must also be a test for a break in the variance, and we did not want to test for such a compound hypothesis. Because the negative binomial specification does not impose this restriction, we ran the negative binomial for all of our basic specifications, and the results were qualitatively similar to what we found with OLS. Breaks in the mean were found in similar locations and were only more statistically significant than those under OLS. We report results from OLS throughout the paper because they are more straightforward to interpret.

[19] We would expect the Andrews critical value to be much higher than the Chow value, given the search for the maximal value of the Wald statistic.

FIGURE 2.—WALD STATISTICS, BY Month



is only 77 monthly observations. In order to produce appropriate small-sample critical values for comparison with our sample test statistics, we generated data using the properties of the actual data, but without a break, and then computed the size of the test for a structural break. In particular, we conditioned on the independent variables in the regression (that is, we ran a separate Monte Carlo analysis for each of the specifications reported in table 2).[20] Monte Carlo results for each specification of independent variables were generated from 10,000 draws under these conditions.[21]

The results of the Monte Carlo analysis of size are reported in table 3, together with the asymptotic critical values for the sup Wald statistic from Andrews (1993). As is to be expected, the critical values for any particular size, regardless of specification, are higher than the conventional Chow critical values.[22] Indeed, the $p = 0.05$ Chow test value actually has a p-value closer to 0.20 when the small sample and window search are considered. The Monte Carlo critical values are fairly consistent across specifications. There is some variation, however, which is not surprising, as the number and nature of independent variables changes.

In comparing our sample test statistics from table 2 with the small-sample critical values in table 3, we find quite strong evidence for the statistical significance of the structural breaks found. For specification A, the sample test statistic is 33.7, which exceeds the critical value at even the 1% level (11.03). For specification B, the sample test statistic is 20.9, which also exceeds critical values beyond the 1% level (11.62). The sample test statistic for specification C is much lower, at 8.89, which exceeds the critical value at the 5% level. As explained earlier, we do not find teen unemployment a convincing explanation for the decline in youth homicide and therefore are doubtful that this decline in the sample test statistic reflects a less significant program effect.

For specifications A and B, the power of the structural break test is not an issue, for we found convincing evidence of structural breaks. However, for specification C, where there is weaker evidence of a break, one might wonder whether this is really less evidence of a break or simply low power of the test. Table 4 reports results of Monte Carlo analysis of power of the test for the same specifications we have considered throughout. Data were generated under the specific alternative hypothesis that there is a break in the mean that matches both the magnitude and the location of the break in the actual data.

Clearly specifications A and B have quite high power to detect breaks. The power for the third specification is quite a bit lower, however, reconfirming our hypothesis that it is harder to find a break, even when one exists, with this set of control variables. In particular, table 4 indicates that for specification C (that is, controlling for population, adult homicide rate, and the questionable youth unemployment rate), we would find a break in mean youth homicide counts only about 43% of the time (using a 5% critical value) even when such a break truly exists. Therefore we should not be surprised that the p-value for specification C is higher than for the other two models.

Overall, the Monte Carlo analysis is quite informative. Because the resulting critical values are higher than the asymptotic values, the need for Monte Carlo analysis when applying the test of structural break is apparent. The Monte Carlo results also shed further light on the properties of the

[20] We conditioned on the independent variables because it is not possible to generate time series data with the properties of the actual data for the right-hand-side variables, because these are measured over too short a time span to accurately determine the time series properties. Conditioning on the independent variables is valid as long as they are exogenous.

[21] In particular, for each specification we ran a Poisson model and computed the variance of the residuals. For size calculations we used the variance over the full period, and for power calculations we used the variance in each of the two subperiods defined by the appropriate breakpoint from the actual data for the specification. For each of 10,000 draws we used these variances to generate 77 observations using the Poisson distribution. For each draw we then ran OLS with the appropriate controls for that specification, calculating the maximum Wald statistic and its location in the time period. The data were generated using the GAUSS Poisson random number generator.

[22] For 77 observations these are: 2.77 at the 10% significance level, 3.97 at 5%, and 6.98 at 1%.

TABLE 3.—MONTE CARLO CRITICAL VALUES: SIZE

	Significance Level				
Model	$p = 0.20$	$p = 0.15$	$p = 0.10$	$p = 0.05$	$p = 0.01$
Asymptotic	—	—	—	6.62	10.26
A	3.70	4.38	5.32	6.98	11.03
B	3.84	4.54	5.49	7.29	11.62
C	4.88	5.64	6.78	8.82	13.76

Monte Carlo critical values result from 10,000 draws of 77 observations from Poisson distribution with variance from the true data, testing for break in mean only. Asymptotic critical values are from Andrews (1993).

TABLE 4.—MONTE CARLO CALCULATIONS: POWER

	Significance Level			
Model	$p = 0.20$	$p = 0.15$	$p = 0.10$	$p = 0.05$
A	99.9%	99.8%	99.6%	99.0%
B	98.2%	97.2%	95.2%	90.5%
C	72.8%	66.8%	57.4%	42.7%

Power is for the associated critical value from Table 3.

test. In particular, the distribution of test statistics is quite nonlinear, with the critical values rising steeply in the upper tail. That is, spurious breaks with high test statistics are found, but rarely.

VI. Robustness

The methodology presented and applied in this paper can tell us whether there is a break in the time series for some outcome variable, and if so, where that break is. Yet there are two types of concerns with regard to the robustness of the findings: the controls may not be adequate, and the specification of the form of the break may not be appropriate. We address four dimensions of these concerns.

A. *Trend in Youth Homicide*

One might be concerned that youth homicide fell not due to the program, but rather due to a secular trend. The first row of table 5 presents the result of expanding specification B to include a linear trend. Adding a trend to this specification reduces the maximum Wald statistic to 6.93 (from 20.93). The month of maximum break is June 1996, which is substantively the same as the results in table 2.[23] The maximum Wald statistic is much lower with the linear trend included. In fact, we would no longer reject the null hypothesis at conventional choices of test size.

To further investigate the effect of including the trend control in the analysis, we considered the possibility of a break in trend, as well as a break in mean. The result of this specification is presented in the second row of table 5. In this case we find that the timing of break is roughly similar (in fact it is identical to specification B and within 2 months of all other specifications). The statistical significance of the break (now a break in two parameters) is also improved from the case of including a trend with only a break in mean. Using the Monte Carlo critical values, the null hypothesis of no break in mean and trend is rejected at the 5% level.

Our interpretation is that including the linear trend absorbed some of the break in mean. Allowing the trend to break reduced the absorption of the mean decline by relaxing the requirement that the trend be linear and suggests that a linear trend was misspecified. In contrast to ordinary regression analysis, in which inclusion of an irrelevant variable does not bias the coefficients, in a structural break test a nuisance variable can be problematic. For example, inclusion of a trend could absorb some of the change in mean, even if there is no trend in the data, that is, if there are just two regimes with different (constant) means.

Overall we conclude that the change in the mean monthly number of youth homicides in Boston was not due to an overall downward trend in the data.[24] By controlling for a general downward trend, as well as the declining trend in adult homicide that occurred at the same time and probably even due to the same program, we are confident that we have ruled out any likely secular trends that might explain the decline in the youth homicide rate.

B. *Break in Adult Homicide*

A related concern for the main results of the paper is the decline in the adult homicide rate (see figure 3). If one believed that adults were a good control group for youth, then it would be natural to allow the adult series to break also, using a difference-in-differences model. Although we have argued that the BGP could be expected to affect the number of older homicide victims through spillovers—so that adults would not be available as a control group—it is nevertheless useful to compare the structural break test with a difference-in-differences approach.

If the number of victims of homicide aged 25 and older is subjected to the same structural break test used above for younger victims (specification A), the sup Wald statistic of 14.05 is located in September of 1996, which is qualitatively similar to the youth result, but slightly later. Restricting one's attention to victims aged 35–44 to reduce the possibility of programmatic spillovers leaves a sup Wald statistic of 4.40 in March of the following year, below the critical value of conventional significance levels. Consistent with the descriptive statistics in table 1 and the discussion of

TABLE 5.—PARAMETER INSTABILITY IN YOUTH HOMICIDE: ROBUSTNESS CHECKS

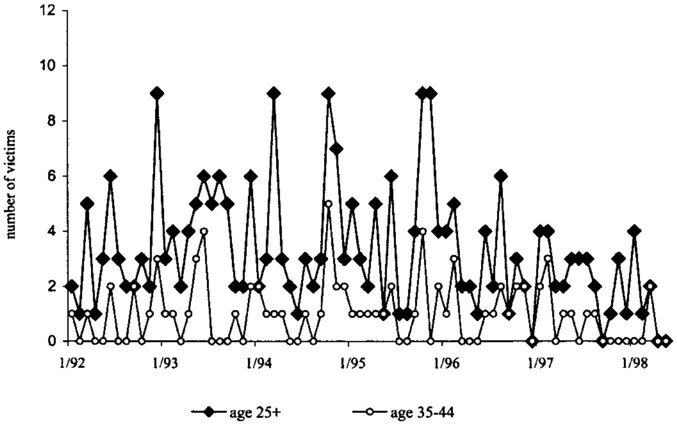
	Maximum		Monte Carlo Critical Values	
Model	Wald Statistic	Month of Max.	$p = 0.10$	$p = 0.05$
B + trend	6.93	June 1996	6.85	8.75
Change in mean and trend	13.24	August 1996	9.88	12.44
Victims 25+	14.05	September 1996	5.19	6.93
Victims 35–44	4.17	March 1997	5.16	6.85

$N = 77$ months, January 1992 through May 1998. Independent variables included in the first two rows are: population, adult homicide rate, trend, and 11 month indicators. Independent variables included in the last two rows are: population and 11 month indicators. Asymptotic critical values were taken from Andrews (1993). Monte Carlo critical values result from 10,000 draws of 77 observations from Poisson distribution with variance from the true data.

[23] The result is slightly different from specification B, which had the maximum break in August 1996 in table 2, but specifications A and C had the maximum break in June 1996, so the general finding is upheld. See figure 2.

[24] We find similar results when we use the national youth homicide rate to represent the counterfactual instead of the linear trend. An analysis of the youth homicide experience of other large U.S. cities over a similar time period revealed no robust or systematic breaks. Together, these results suggest the Boston experience was unique.

FIGURE 3.—MONTHLY ADULT HOMICIDE COUNT 1/92–5/98



spillovers in section II, the effect sizes associated with these breaks are smaller than for the youth homicide series (50% declines versus 70% declines).

In addition we ran a difference-in-differences model using the break date found in the youth homicide series alone. Using all victims aged 25 and older as the control group, youth homicide dropped by approximately 1 victim per month relative to the decline for adults. That is, in the summer of 1996, both groups had sharp 40% drops relative to the preprogram period (p-value 0.001), and youth dropped an additional 45% relative to adults (p-value 0.03). We find that the number of adult victims declined, in addition to the number of youth victims, but that the effect on youth is greater. Overall, both the timing and magnitude of the youth and adult effects are consistent with the spillovers described in section II.

C. *Searching Outside the Window*

By restricting the search to a window of possible program leads or lags one runs the risk of misspecifying the time series if there are breaks outside the window. For example, a break in the series in the opposite direction of a program effect and outside the window could lead a researcher to erroneously conclude there was no break. More germane to this setting, if an earlier program had led to a substantial decline in homicide shortly before January 1996, and we only searched for a break following January 1996, we could conclude there was a very large break around the time of our program when in fact it had a small effect. To check for this possible threat to robustness we searched the entire time series for a possible break. We found that the sup Wald statistic for the entire time series was located in the window.

D. *Gradual Program Effect*

It is possible to have a program that suggests a more complicated program effect than a one-time change in mean. If one thought there were going to be a gradual implementation, a break in mean and trend could be allowed. In our case, the change in mean and trend is characterized by a large immediate decline and change in trend in the positive direction. The change in trend is small, especially relative to the substantial decline in the mean number of homicides, and the program effect is dominated by a change in mean.

Alternatively, one might think of implementation in stages, allowing multiple points of impact, in which case multiple changes in mean could be allowed. Following Bai and Perron (1998), we allow for a second break in specification B, conditioning on the first break. The sup Wald statistic for the second break is only 5.42 in November 1996. This value is substantially below the 5% asymptotic critical value of 11.1, and therefore the null hypothesis of no second break, conditional on the first, is not rejected.[25] Considering the results on the multiple break and on the break in mean and trend together, it does not appear that the decline in youth homicide in Boston took place in discrete stages or as a gradual decline.

Overall, these results confirm the robustness of the conclusion that there was a structural break in youth homicide in Boston in the summer of 1996. This break is not easily explained by secular trends, nor is it the erroneous attribution of an earlier break in the time series. Finally, the form of the break is best characterized as a discrete shift in mean, not a gradual or multistep transition.

VII. Conclusion

Substantively, this evaluation has found that there was a statistically significant break in mean associated with a substantial decrease in youth homicide in the summer of 1996. This discontinuity coincides with when the BGP was implemented. Controlling for population, the adult homicide rate, and a linear trend, we have confidence that we have captured a program effect rather than an unrelated change in youth homicide. Any alternative explanation for the drop in youth homicide over this period would need to be able to account for the suddenness of the change at that time, and we have not discovered any convincing explanations with this quality.

While our application tested for a break in mean because the model underlying the intervention was a tipping one, many other types of effects can be evaluated with this procedure. One can test for a break in the entire set of controls (regime shift) or, as more likely relevant to program evaluation, for a break in trend or a break in the relationship of the outcome to a single control variable. The structural break test is useful because it can identify timing and statistical significance of program effects even when the timing of effect is uncertain a priori, and can give different inference from the usual methods. The last point is not a technical detail. Traditional Chow tests overstate statistical

[25] We did not do a Monte Carlo analysis of multiple breaks, due to the small sample size.

significance when used for program evaluation. Given that the primary motivation for evaluating a program is to test whether an intervention "worked," using appropriate methods for statistical inference is essential.

REFERENCES

Andrews, Donald W. K., "Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica* 61:4 (1993), 821–856.

Andrews, Donald W. K., and Werner Ploberger, "Optimal Tests When a Nuisance Parameter is Present Only under the Alternative," *Econometrica* 62:6 (1994), 1383–1414.

Bai, Jushan, "Estimating Multiple Breaks One at a Time," *Econometric Theory* 13 (1997), 315–352.

Bai, Jushan, and Pierre Perron, "Estimating and Testing Linear Models with Multiple Structural Changes," *Econometrica* 66:1 (1998), 47–78.

Banerjee, Anindya, Robin L. Lumsdaine, and James H. Stock, "Recursive and Sequential Tests of the Unit Root and Trend Break Hypotheses: Theory and International Evidence," *Journal of Business & Economic Statistics* 10 (1992), 271–287.

Brown, R. L., J. Durbin, and J. M. Evans, "Techniques for Testing the Constancy of Regression Relationships over Time with Comments," *Journal of the Royal Statistical Society, Series B*, 37 (1975), 149–192.

Cook, Philip J., and John H. Laub, "The Epidemic in Youth Violence" (pp. 27–64), in Michael Tonry and Mark H. Moore (Eds.), *Youth Violence* (1998).

Kennedy, David M., Anne M. Piehl, and Anthony A. Braga, "Youth Violence in Boston: Gun Markets, Serious Youth Offenders, and a Use-Reduction Strategy," *Law and Contemporary Problems* 59:1 (1996), 147–183.

Levitt, Steven D., and Sudhir Alladi Venkatesh, "An Economic Analysis of a Drug-Selling Gang's Finances," *Quarterly Journal of Economics* 115 (2000), 755–789.

Nyblom, Jukka, "Testing for the Constancy of Parameters over Time," *Journal of the American Statistical Association* 84 (1989), 545–549.

Piehl, Anne Morrison, "Economic Conditions, Work, and Crime" (pp. 302–319), in Michael Tonry (Ed.), *Handbook on Crime and Punishment* (Oxford University Press, 1998).

Piehl, Anne Morrison, David M. Kennedy, and Anthony A. Braga, "Problem Solving and Youth Violence: An Evaluation of the Boston Gun Project," *American Law and Economics Review* 2:1 (2000), 58–106.

Quandt, Richard E. "Tests of the Hypothesis That a Linear Regression System Obeys Two Separate Regimes," *Journal of the American Statistical Association* 55 (1960), 324–330.

Stock, James H., "Unit Roots, Structural Breaks and Trends" (pp. 2740–2841), in R. F. Engle and D. L. McFadden (Eds.), *Handbook of Econometrics.* vol. 4 (New York: Elsevier Science, 1994).